Supplément au Chapitre 3

3.9 : Arbres de régression multivariable ou « multivariate regression trees »

3.9.1 Introduction

L'arbre de régression multivariable (ARM) est une généralisation de l'arbre univariable CART de Breiman (1984) qui a fait sa première apparition dans la littérature dans les années 90 (De'ath, 2002 ; Larsen, 2004 ; Segal, 1992). C'est un type de groupement hiérarchique divisif tel celui de Ward. On commence donc la construction du groupement avec tous les objets dans un même groupe, puis on divise en sous-groupes tant que chacun des objets ne forme pas un groupe. Le critère à minimiser pendant la construction est également la somme des carrés intra-groupes, i.e. le même que le groupement de Ward. En revanche, trois particularités importantes (l'une découlant de l'autre) distinguent l'ARM de la méthode de Ward :

1) L'arbre de régression multivariable est un type de groupement sous contrainte : en plus du critère à minimiser (SC intra-groupe), se calculant sur une matrice dite « réponse », on impose une contrainte supplémentaire; cette dernière est une matrice de variables explicatives. Par conséquent, les groupes sont délimités par ces variables. On illustrera ce principe plus loin.

2) Cette première caractéristique lui confère l'avantage suivant : puisque les groupes sont caractérisés par des variables explicatives, il est possible de faire des prédictions de la réponse sur de nouveaux objets pour lesquels on détient les valeurs des variables explicatives.

3) Puisqu'il est possible de faire des prévisions sur de nouveaux objets, il est également à notre portée de trouver une partition optimale en termes de prédictions. On utilise communément la validation croisée avec v-recouvrements (le plus souvent $v = 10$) afin d'arriver à nos fins. Des détails complémentaires sur ces différentes procédures sont fournis dans les sections à venir.

Plusieurs chercheurs ont par le passé utilisé l'ARM à des fins d'exploration de données écologiques (De'ath, 2002 ; Larsen, 2004 ; Ouellette, 2005). La plupart des exemples présents dans la littérature se limitent à la prédiction d'assemblages d'espèces à de nouveaux sites dont on connaît les variables explicatives d'intérêt (physico-chimiques par exemple). Puisqu'il s'agit d'une méthode non-paramétrique, elle s'applique donc particulièrement bien à des données qui ne sont pas distribuées normalement. De plus, le processus interne nous assure simultanément la

construction du modèle et la sélection des variables explicatives d'intérêt, et par le fait même s'adapte remarquablement bien à un grand nombre de variables explicatives quantitatives et/ou qualitatives.

3.9.2 Construction du modèle

Nous allons maintenant nous attarder à la présentation des étapes de la méthode de l'ARM qui consistera en deux volets : la construction de l'arbre, et la sélection de la partition finale à l'aide de la validation croisée.

Premier volet : Construction de l'arbre

Comme mentionné plus haut, l'ARM permet de faire l'analyse de deux tableaux, soit ceux des variables explicatives (**X**) et de la matrice réponse (**Y**) comme en témoigne la Figure 1. On supposera, pour alléger le texte, que **Y** est une matrice d'abondances d'espèces, et que **X** est une matrice de caractéristiques physico-chimiques des mêmes sites. Afin de simplifier l'explication de la méthode, nous nous limiterons au cas de deux variables explicatives. On généralisera par la suite.

Y		
1	0	0
5	0	0
4	4	0
7	8	1
2	3	8
2	5	2
0	0	1
0	1	0
0	0	0
1	1	1
2	2	3
3	3	3

X_1	X_2
1	0
9	1
3	3
4	7
10	8
6	9
7	2
8	3
9	4
1	3
3	4
4	5

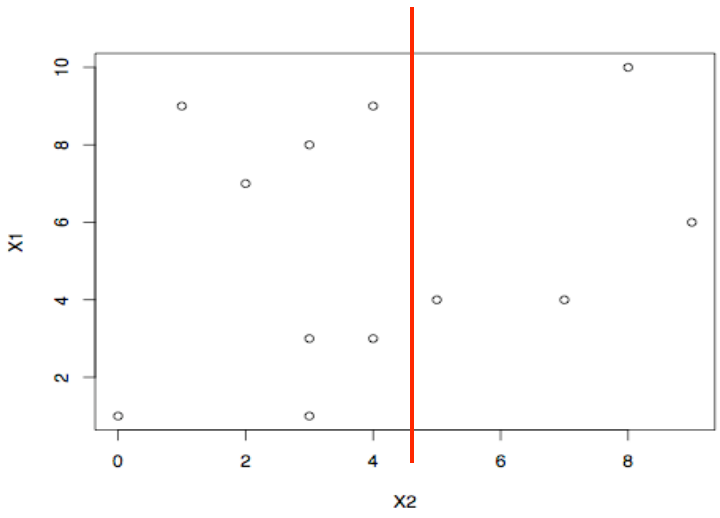

Figure 1 : Illustration d'un exemple de construction d'ARM. **Y** est une matrice réponse de composition spécifique, et **X** une matrice de variables explicatives comprenant deux variables X_1 et X_2. On a, à droite, le diagramme de dispersion des objets dans l'espace des variables explicatives. Le trait rouge représente la première bipartition. Les deux groupes de sites sont représentés dans la matrice de variables explicatives **X** par un code de couleur : gris $X_2 <= 4.5$, blanc $X_2 > 4.5$.

Comme le montre le diagramme de dispersion, le modèle est construit dans l'espace des variables explicatives, à partir de bipartitions consécutives. En d'autres mots, on divise plusieurs fois les objets en deux groupes (la figure 1 montre la première bipartition en rouge).
On aurait très bien pu faire la première bipartition à $X_2 = 2$ au lieu de $X_2 = 4{,}5$.
Comment a-t-on fait ce choix ?

On commence par examiner, pour chaque variable, chacune des bipartitions possibles qui crée deux groupes de sites.

On calcule par la suite, pour chacune de ces bipartitions, la somme des carrés des écarts à la moyenne du groupe de la matrice réponse selon la formule ci-dessous. Celle dont la valeur est minimale l'emporte.

Formule 1

$$\sum_{k=1}^{g}\sum_{i=1}^{n_k}\sum_{j=1}^{3}\left(y_{ij} - \overline{y}_{jk}\right)^2$$

où k est le nombre de groupes, i est le nombre d'éléments dans le groupe k, j correspond à l'espèce (nous avons trois espèces dans cet exemple), $\overline{y}_{jk}$ est la moyenne pour l'espèce j dans le groupe k, et y_{ij} est l'abondance de l'espèce j dans le site i pour chaque groupe k.

La bipartition finale nous procure deux éléments d'une importance capitale:

1) Les deux nouveaux groupes selon cette bipartition.

2) La variable explicative ainsi que le seuil permettant de bipartitionner les sites. C'est ce qui différencie l'arbre de régression multivariable d'une méthode de groupement comme Ward ou *k*-means. Cela nous permet, une fois le modèle construit, de faire des prédictions d'assemblages d'espèces à de nouveaux sites dès que les valeurs des variables explicatives sont disponibles.

Une fois la première bipartition accomplie, on recommence cette étape avec les deux nouveaux groupes et ainsi de suite. Pour les deux variables, on poursuit par le calcul de la somme des carrés des écarts à la moyenne du groupe, et ce pour toutes les bipartitions de tous les groupes. Le processus doit continuer tant qu'un groupe contient plus d'un site. À la figure 2, nous n'avons pas effectué toutes les bipartitions, mais la partie de droite en montre plusieurs. La partie de gauche quant à elle présente les résultats sous forme d'arbre.

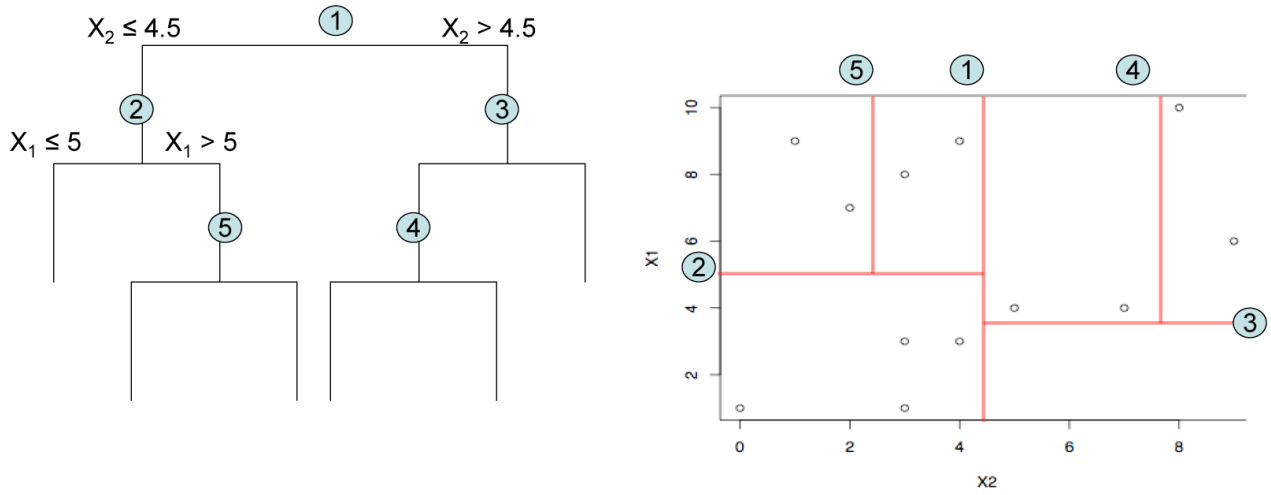

Figure 2 : À droite, plusieurs bipartitions effectuées lors de la construction de l'ARM; à gauche, représentation de ces bipartitions sous forme d'arbre. Les données utilisées sont celles de la figure 1.

Ceci clôt le premier volet de la construction du modèle. L'autre partie est indispensable à l'obtention du modèle final. Cependant, ne pas avoir recours à une méthode de sélection de modèle impliquerait l'obtention d'un arbre avec autant de groupes que de sites, ce qui perdrait toute utilité. C'est pour cela qu'il est pertinent d'utiliser une méthode d'élagage de l'arbre ; nous suggérons la validation croisée (possiblement à *v*-recouvrements).

Deuxième volet : la validation croisée

La validation croisée fait partie intégrante du processus de construction de l'arbre. La première étape débute avant même la première bipartition. On divise l'ensemble des objets du noeud racine en v (souvent $v = 10$) groupes composés, autant que possible, du même nombre de sites. Afin d'illustrer le processus dans la figure 3, seulement 5 recouvrements ont été réalisés.

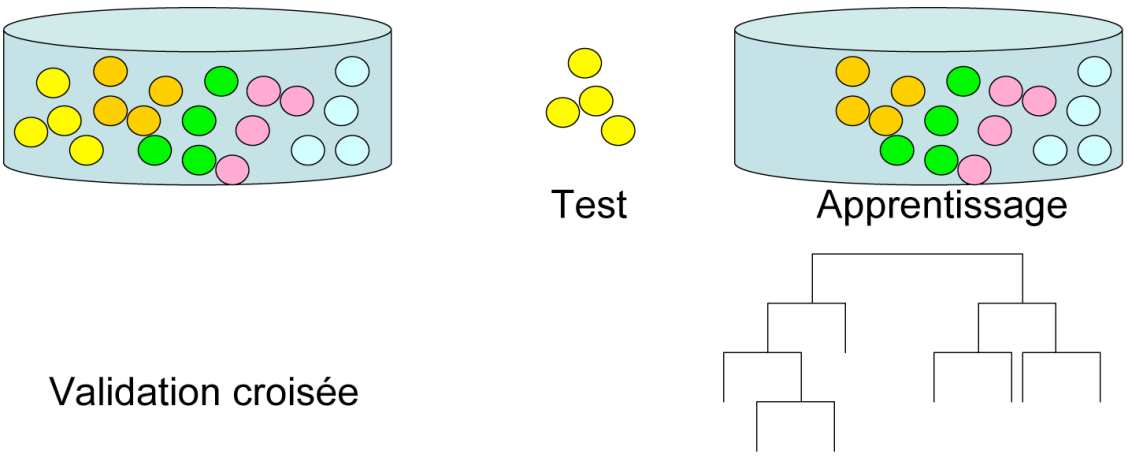

Figure 3 : Illustration du processus de la validation croisée à *v*-recouvrements. Cette figure montre 5 recouvrements.

L'étape suivante consiste à retirer un groupe du jeu de données, l'un à la suite de l'autre et un seul à la fois pour ensuite construire le modèle en question à partir des objets restants. Nous allons distinguer entre le groupe d'objets retirés, appelé le *groupe test*, et les objets restants que l'on nomme le *groupe d'apprentissage*.

Une fois le modèle construit à l'aide du groupe d'apprentissage (le modèle 1 de la figure 4), on utilise les objets du groupe test (représentés en jaune) afin de calculer l'erreur relative de la validation croisée, qui consiste en la somme des carrés des écarts entre la valeur prédite et la valeur réelle, divisée par la somme des carrés des écarts autour de la moyenne générale correspondant aux objets de l'ensemble test.

Plus précisément, en se référant à la formule de la figure 4, si y_{ij} représente la valeur réelle, « y chapeau » sera l'abondance estimée par le modèle, « y barre » représentera la moyenne sachant que n_p est le nombre d'objets dans l'ensemble test et finalement m est le nombre d'espèces dans la matrice réponse.

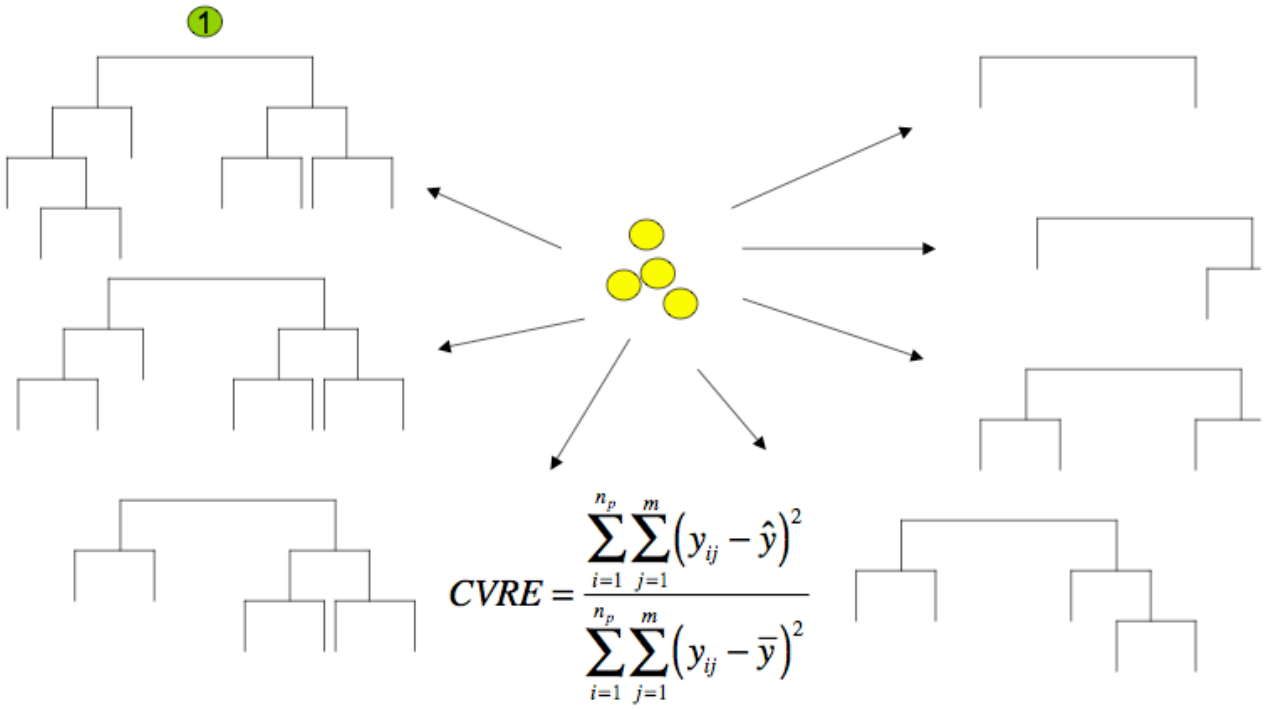

Figure 4 : Illustration du principe de la validation croisée à *v*-recouvrements et de la formule de l'erreur relative associée.

L'erreur relative de validation croisée peut varier de 0, pour un bon modèle prévisionel, à des valeurs proches de 1, pour un modèle avec pauvre potentiel prévisionnel. Cette valeur sera calculée pour tous les modèles imbriqués du modèle 1, tel qu'illustré dans la figure 4. Le processus sera répété un certain nombre de fois (disons 500) et le modèle final choisi sera celui qui présente la plus petite erreur relative de validation croisée.

3.9.3 Généralisation : Hypervolumes rectangulaires des variables explicatives

Revenons maintenant à la construction du modèle, que l'on peut facilement généraliser à plusieurs variables. On s'est jusqu'à maintenant limité au cas à deux variables explicatives. Si l'on revient à la figure 2, chacun des assemblages est caractérisé par un rectangle dont chacun des côtés est délimité par les valeurs de seuils associés aux variables explicatives qui le caractérisent. L'ajout d'une dimension

implique que les assemblages sont délimités par des boîtes rectangulaires. En poursuivant ce processus de généralisation, les assemblages seront, à chaque nouvelle dimension ajoutée, délimités par des hypervolumes physico-chimiques.

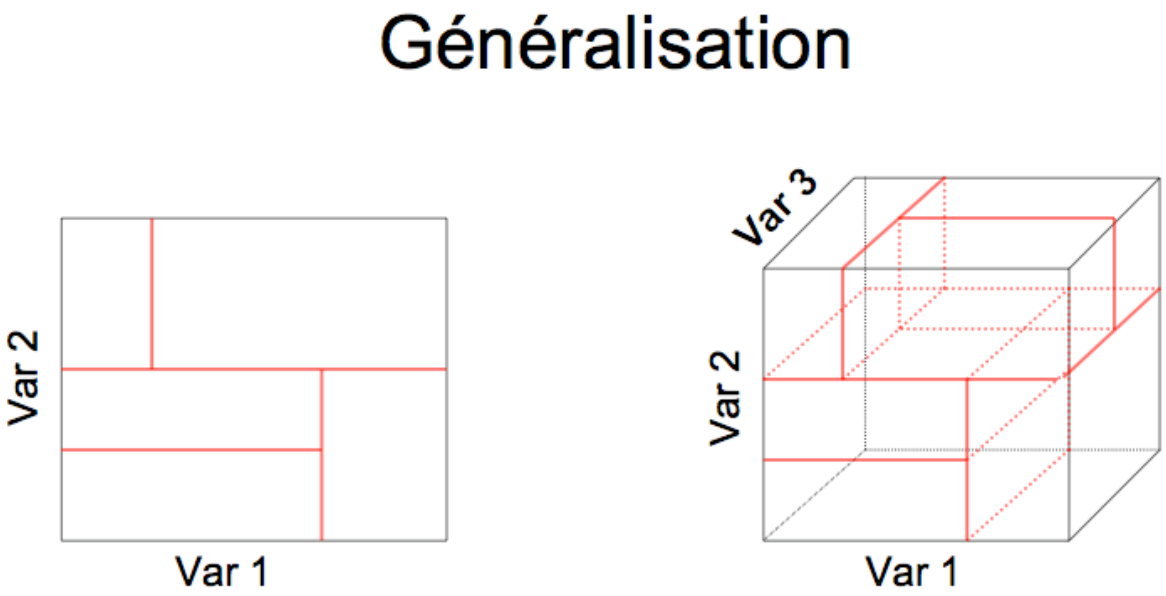

Figure 5 : Généralisation du processus de construction de l'ARM de la figure 2.

Cela vous rappelle-t-il quelque chose ? Le modèle final nous procure un résultat qui est ***analogue*** à la théorie de la niche telle que décrite par Hutchinson (1957). La seule différence étant que dans notre cas, l'hypervolume physico-chimique est délimité par un assemblage d'espèces, et non par seulement une espèce.

3.9.4 La stratégie invoquée par l'ARM

Quand vient le temps de faire le lien entre des données spécifiques et les variables environnementales associées, les stratégies adoptées être de plusieurs types. Nous en avons retenu deux des plus appropriées.
1) On pourrait utiliser une approche permettant au biote de « raconter son histoire » avant d'en déduire les liens avec les variables environnementales spécifiques (Clarke,

1993). Par exemple, si les variables physico-chimiques sont toutes quantitatives, il sera tout désigné de projeter les résultats d'un groupement *k*-means ou Ward (construit sur les données d'abondance d'espèces) sur une carte géographique avec des courbes de contour (montrant les variables environnementales), ou encore dans un espace réduit des variables environnementales tiré d'une ACP (analyse en composantes principales, voir Chapitre 4a pour plus de détails).

2) Une autre option s'offre à nous : elle repose sur la supposition nécessaire que des sites similaires en termes de variables physico-chimiques le soient aussi en termes de composition spécifique; cela suppose que les variables explicatives d'intérêt sont incluses dans l'analyse (Clarke, 1993), ce qui est le cas de l'ARM (lorsque les variables explicatives ont été bien choisies).

3.9.5 Le bruit et l'ARM

La plupart des jeux de données contiennent du bruit (variation aléatoire). Cela peut embrouiller le partitionnement en faisant augmenter la somme des carrés des écarts et par conséquent produire un arbre contenant moins de groupes, un phénomène indésirable, particulièrement dans le cas qui nous concerne. Une solution possible à ce problème consisterait à effectuer une ordination des sites par rapport aux variables environnementales (voir chapitre 4a) et en utiliser les vecteurs propres comme variables explicatives. On devrait normalement obtenir une partition contenant plus de groupes. Il est important de noter qu'il sera toujours possible de faire des prévisions : il suffira de calculer les positions des nouveaux sites dans l'espace réduit (voir Legendre et Legendre 1998 pour plus de détails), et d'utiliser les scores sur chacun des vecteurs propres de chaque nouveau site afin de prévoir son assemblage d'espèces.

3.9.6 Interprétation des résultats

3.9.6.1 Interprétation des seuils des variables explicatives choisies

L'interprétation du modèle d'ARM est souvent qualifiée de simple. Il faut tout de même prêter une attention particulière à la manière d'interpréter les seuils issus du programme utilisé. Revenons à la figure 2; si l'on examine plus attentivement le diagramme de dispersion, on remarque que le seuil de la première bipartition optimale est $X_2 = 4.5$. On aurait pourtant pu utiliser 4.1 , 4.55 ou encore 4.999. La valeur de 4.5 fournie par le programme est en réalité la moyenne entre les bornes 4 et 5 qui se trouvent dans le jeu de données. Comme nous nous trouvons dans une zone d'incertitude (puisqu'on n'a pas de valeurs de X_2 entre 4 et 5), on donne par défaut la moyenne de ces valeurs.

3.9.6.2 Les espèces discriminantes

À un nœud donné, nous désignerons par le terme "espèces discriminantes" celles qui contribuent le plus au coefficient de détermination, ou le moins à l'erreur relative. Par définition, toute espèce qui aura une abondance très différente d'un groupe à un autre au sein d'une même bipartition entre dans cette catégorie. On peut alors émettre l'hypothèse que cette espèce répond fortement à la variable explicative utilisée pour faire la bipartition (une relation de corrélation et non une relation de causalité). C'est Dea'th (2002) qui a introduit pour la première fois cette notion dans son article. Malheureusement, on ne peut obtenir ce résultat intermédiaire directement avec la fonction 'mvpart' que l'on verra plus en profondeur dans la section pratique. Pour parer à ce problème, on doit faire les calculs manuellement.

Références

Breiman, L., J. H. Friedman, et al. (1984). Classification and Regression Trees. Belmont, California, USA, Wadsworth International Group.

Clarke, K. R. and M. Ainsworth (1993). "A method of linking multivariate community structure to environmental variables." Marine Ecology Progress Series **92**(3): 205-219.

De'ath, G. (2002). "Multivariate regression trees : a new technique for modeling species-environment relationships." Ecology **83**(4): 1105–1117.

Larsen, D. R. and P. L. Speckman (2004). "Multivariate Regression Trees for Analysis of Abundance Data." Biometrics **60**: 543-549.

Legendre, P. and L. Legendre (1998). Numerical Ecology, Elservier.

Ouellette, M.-H., J.-L. DesGranges, et al. (2005). L'arbre de régression multivariables: classification d'assemblage d'oiseaux fondée sur les caractéristiques de leur habitat. Société Francophone de Classification, Montréal.